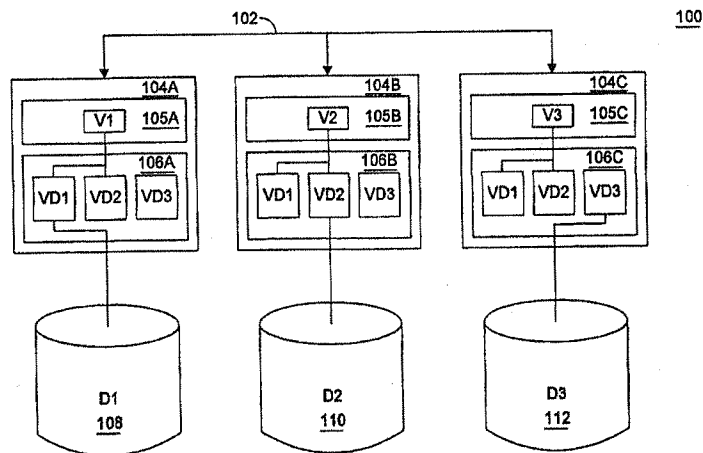




INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G06F 11/14		A1	(11) International Publication Number: WO 99/59064
			(43) International Publication Date: 18 November 1999 (18.11.99)
(21) International Application Number: PCT/US99/09903		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).	
(22) International Filing Date: 7 May 1999 (07.05.99)			
(30) Priority Data: 09/076,388 12 May 1998 (12.05.98) US 09/076,347 12 May 1998 (12.05.98) US 09/076,346 12 May 1998 (12.05.98) US 09/076,274 12 May 1998 (12.05.98) US			
(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, Palo Alto, CA 94303 (US).			
(72) Inventors: SLAUGHTER, Gregory, L.; 3326 Emerson Street, Palo Alto, CA 94306 (US). TRAVERSAT, Bernard, A.; Apartment 402, 2055 California Street, San Francisco, CA 94109 (US). HERNDON, Robert; 1837 S. Nevada Avenue #137, Colorado Springs, CO 80906 (US). ZHENG, Xiaoyan; 5454 Golubin Common, Fremont, CA 94555 (US). BLOCK, Robert, J.; 265 N. Rengstorff #29, Mountain View, CA 94043 (US).		Published With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.	
(74) Agent: KIVLIN, B., Noël; Conley, Rose & Tayon, P.C., P.O. Box 398, Austin, TX 78767-0398 (US).			

(54) Title: HIGHLY AVAILABLE CLUSTER VIRTUAL DISK SYSTEM



(57) Abstract

A cluster implements a virtual disk system that provides each node of the cluster access to each storage device of the cluster. The virtual disk system provides high availability such that a storage device may be accessed and data access requests are reliably completed even in the presence of a failure. To ensure consistent mapping and file permission data among the nodes, data are stored in a highly available cluster database. Because the cluster database provides consistent data to the nodes even in the presence of a failure, each node will have consistent mapping and file permission data. A cluster transport interface is provided that establishes links between the nodes and manages the links. Messages received by the cluster transports interface are conveyed to the destination node via one or more links. The configuration of a cluster may be modified during operation. Prior to modifying the configuration, a reconfiguration procedure suspends data access requests and waits for pending data access requests to complete. The reconfiguration is performed and the mapping is modified to reflect the new configuration. The node then updates the internal representation of the mapping and resumes issuing data access requests.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

TITLE: **HIGHLY AVAILABLE CLUSTER VIRTUAL DISK SYSTEM****BACKGROUND OF THE INVENTION**5 1. Field of the Invention

This invention relates to the field of distributed computing systems and, more particularly, to distributed virtual storage devices.

10 2. Description of the Related Art

Distributed computing systems, such as clusters, may include two or more nodes, which may be employed to perform a computing task. Generally speaking, a node is a group of circuitry designed to perform one or more computing tasks. A node may include one or more processors, a memory and interface circuitry. Generally speaking, a cluster is a group of two or more nodes that have the capability of exchanging data between nodes. A particular computing task may be performed upon one node, while other nodes perform unrelated computing tasks. 15 Alternatively, components of a particular computing task may be distributed among the nodes to decrease the time required perform the computing task as a whole. Generally speaking, a processor is a device configured to perform an operation upon one more operands to produce a result. The operations may be performed in response to instructions executed by the processor.

Nodes within a cluster may have one or more storage devices coupled to the nodes. Generally speaking, a 20 storage device is a persistent device capable of storing large amounts of data. For example, a storage device may be a magnetic storage device such as a disk device, or optical storage device such as a compact disc device. Although a disk device is only one example of a storage device, the term "disk" may be used interchangeably with "storage device" throughout this specification. Nodes physically connected to a storage device may access the storage device directly. A storage device may be physically connected to one or more nodes of a cluster, but the 25 storage device may not be physically connected to all the nodes of a cluster. The nodes which are not physically connected to a storage device may not access that storage device directly. In some clusters, a node not physically connected to a storage device may indirectly access the storage device via a data communication link connecting the nodes.

It may be advantageous to allow a node to access any storage device within a cluster as if the storage 30 device is physically connected to the node. For example, some applications, such as the Oracle Parallel Server, may require all storage devices in a cluster to be accessed via normal storage device semantics, *e.g.*, Unix device semantics. The storage devices that are not physically connected to a node, but which appear to be physically connected to a node, are called virtual devices, or virtual disks. Generally speaking, a distributed virtual disk system is a software program operating on two or more nodes which provides an interface between a client and one 35 or more storage devices, and presents the appearance that the one or more storage devices are directly connected to the nodes. Generally speaking, a client is a program or subroutine that accesses a program to initiate an action. A client may be an application program or an operating system subroutine.

Unfortunately, conventional virtual disk systems do not guarantee a consistent virtual disk mapping. Generally speaking, a storage device mapping identifies to which nodes a storage device is physically connected 40 and which disk device on those nodes corresponds to the storage device. The node and disk device that map a

virtual device to a storage device may be referred to as a node/disk pair. The virtual device mapping may also contain permissions and other information. It is desirable that the mapping is persistent in the event of failures, such as a node failure. A node is physically connected to a device if it can communicate with the device without the assistance of other nodes.

5 A cluster may implement a volume manager. A volume manager is a tool for managing the storage resources of the cluster. For example, a volume manager may mirror two storage devices to create one highly available volume. In another embodiment, a volume manager may implement striping, which is storing portions of files across multiple storage devices. Conventional virtual disk systems cannot support a volume manager layered either above or below the storage devices.

10 Other desirable features include high availability of data access requests such that data access requests are reliably performed in the presence of failures, such as a node failure or a storage device path failure. Generally speaking, a storage device path is a direct connection from a node to a storage device. Generally speaking, a data access request is a request to a storage device to read or write data.

 In a virtual disk system, multiple nodes may have representations of a storage device. Unfortunately, 15 conventional systems do not provide a reliable means of ensuring that the representations on each node have consistent permission data. Generally speaking, permission data identify which users have permission to access devices, directories or files. Permissions may include read permission, write permission or execute permission.

 Still further, it is desirable to have the capability of adding or removing nodes from a cluster or to change the connection of existing nodes to storage devices while the cluster is operating. This capability is particularly 20 important in clusters used in critical applications in which the cluster cannot be brought down. This capability allows physical resources (such as nodes and storage devices) to be added to the system, or repair and replacement to be accomplished without compromising data access requests within the cluster.

SUMMARY OF THE INVENTION

25 The problems outlined above are in large part solved by a highly available virtual disk system in accordance with the present invention. In one embodiment, the highly available virtual disk system provides an interface between each storage device and each node in the cluster. From the node's perspective, it appears that each storage device is physically connected to the node. If a node is physically connected to a storage device, the virtual disk system directly accesses the storage device. Alternatively, if the node is not physically connected to a 30 storage device, the virtual disk system accesses the storage device through another node in the cluster that is physically connected to the storage device. In one embodiment, the nodes communicate through a data communication link. Whether a storage device is directly accessed or accessed via another node is transparent to the client accessing the storage device.

 In one embodiment, the nodes store a mapping of virtual disks to storage devices. For example, each 35 active node may store a mapping identifying a primary node/disk pair and a secondary node/disk pair for each virtual device. Each node/disk pair identifies a node physically coupled to the storage device and a disk device on that node that corresponds to the storage device. The secondary node/disk pair may also be referred to as an alternate node/disk pair. If the node is unable to access a storage device via the primary node/disk pair, the node may retry the data access request via the secondary node/disk pair. To maintain a consistent mapping between the

nodes in the presence of failures, the mapping may be stored in a highly available database. Because the highly available database maintains one consistent copy of data even in the presence of a failure, each node that queries the highly available database will get the same mapping. The highly available database may also be used to store permission data to control access to virtual devices. Because the highly available database maintains one consistent copy of permission data even in the presence of a failure, each node that queries the database will get the same permission data.

One feature of a virtual disk system in accordance with the present invention is the high availability of the system. In one embodiment, the virtual disk system stores all of the data access requests it receives and retries those requests if an error occurs. For example, the virtual disk system of a node that initiates a data access request, called a requesting node, may store all outstanding data requests. If the destination node, i.e. the node to which the data access request is directed, is unable to complete the data access request, an error indication may be returned to the requesting node and the requesting node may resend the data access request to an alternate node that is connected to the storage device. This error detection and retry is performed automatically and is transparent to the client. In another example, if a node failure occurs, the virtual disk system may receive a modified list of active nodes and resend incomplete data access requests to active nodes coupled to the storage device. This reconfiguration and retry also is transparent to the client.

Another feature of a virtual disk system in accordance with the present invention is the ability to reconfigure the cluster while the cluster is operating. When a cluster is reconfigured, the mapping of virtual disks to storage devices may be updated. To prevent errors, a synchronization command may be performed or operated to all the nodes of the cluster prior to updating the mapping. The synchronization command causes the nodes to stop issuing data access requests. After the mapping is updated, another synchronization command causes the node to resume issuing data access requests.

The virtual disk system may be designed to serve as an interface between a volume manager and storage devices or between a client and a volume manager. In the former configuration, the client interfaces to the volume manager and the volume manager interfaces to the virtual disk system. In the latter configuration, the client interfaces to the virtual disk system and the virtual disk system interfaces to the volume manager.

BRIEF DESCRIPTION OF THE DRAWINGS

Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which:

Figure 1 is a block diagram of a cluster configuration according to one embodiment of the present invention.

Figure 2 is a block diagram of an alternative cluster configuration according to one embodiment of the present invention.

Figure 3 is a block diagram of a virtual disk system operating on two nodes of a cluster according to one embodiment of the present invention.

Figure 4 is a block diagram illustrating the initialization of a netdisk driver according to one embodiment of the present invention.

Figure 5 is a block diagram illustrating the initialization of a cluster transport interface according to one embodiment of the present invention.

Figure 6 is a flowchart diagram illustrating the operation of a virtual disk system according to one embodiment of the present invention.

5 Figure 7 is a flowchart diagram illustrating the initiation of a netdisk driver according to one embodiment of the present invention.

Figure 8 is a flowchart diagram illustrating the initiation of a cluster transport interface according to one embodiment of the present invention.

10 Figure 9 is a block diagram of a cluster transport interface according to one embodiment of the present invention.

Figure 10 is a diagram illustrating permission data according to one embodiment of the present invention.

Figure 11 is a flowchart diagram illustrating the storage and access of consistent permission data according to one embodiment of the present invention.

15 Figure 12 is a flowchart diagram illustrating the update of a configuration mapping according to one embodiment of the present invention.

While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that the drawings and detailed description thereto are not intended to limit the invention to 20 the particular form disclosed, but on the contrary, the intention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DETAILED DESCRIPTION OF THE INVENTION

Turning now Figure 1, a block diagram of a cluster configuration according to one embodiment of the 25 present invention is shown. Cluster 100 includes a data communication link 102, three nodes 104A-104C, and three storage devices 108, 110 and 112. Data communication link 102 provides a data communication path for transferring data between the nodes. Data communication link 102 contemplates a multi-drop link or point-to-point links. For example, data communication link 102 may include three point-to-point links. A first link may provide a communication path between nodes 104A and 104B, a second link may provide a communication path between 30 nodes 104A and 104C, and a third link may provide a communication path between nodes 104B and 104C. In one embodiment, data communication link 102 implements a scalable coherent interface (SCI). In one particular embodiment, the cluster implements a TCP/IP protocol for transferring data over the SCI. It is noted that three nodes are shown for illustrative purposes only. Other embodiments may employ more or less nodes.

35 In the illustrating embodiment, storage device 108 is physically connected to node 104A, storage device 110 is physically connected to node 104B and storage device 112 is physically connected to node 104C. Storage devices 108-112 typically have storage capacities that exceed the storage capacities of the memory of the nodes to which they are connected. Data may be stored in storage devices 108-112 which is not currently being used by a node, and data from the storage device may be stored, or cached, in the memory of the node when the data is needed. In the illustrated embodiment, the storage devices are physically connected to only one node. In

alternative embodiments, a storage device may be physically connected to a plurality of nodes. Multiple physical connections allow a storage device to be accessed even if one node physically connected to the device fails or a storage device path fails.

Multiple instances of the same distributed program may operate on each node. For example, volume manager 105A and volume manager 105B are different instances of the same distributed volume manager program. These instances may communicate with each other via data communication link 102. Each instance is given the same reference number followed by a unique letter, *e.g.*, 105A or 105B. For simplicity, the distributed program may be referred to collectively using only the reference number, *e.g.*, volume manager 105.

Node 104A includes a volume manager 105A and a virtual disk system 106A. In the illustrated embodiment, virtual disk system 106A provides an interface between volume manager 105 and storage devices 108-112. From the perspective of volume manager 105A, each storage device appears to be physically connected to node 104A. Virtual disk system 106 is a distributed program operating on a plurality of nodes. In the illustrated embodiment, an instance of virtual disk system 106 is operating on each node. Virtual disk system 106A, which is the instance of virtual disk system 106 operating on node 104A, includes three virtual devices (VD1, VD2 and VD3) that represent storage devices 108-112, respectively. Volume manager 105 communicates to the virtual devices in the same manner that it communicates to storage devices physically connected to the node. In one embodiment, volume manager 105 uses Unix device driver semantics. Data access requests to storage device 108 (*i.e.* VD1) are conveyed from virtual disk system 106A directly to storage device 108. Data access requests to storage devices 110 and 112 (*i.e.* VD2 and VD3) are conveyed over data communication link 102 to the respective nodes physically connected to those devices.

It is noted that the virtual disks on each node are distinct devices. For example, VD1 on nodes 104A, 104B and 104C are each a unique device managed by a unique device driver. Although the devices are unique, each VD1 device maps to the same physical storage device. In other words, writing data to VD1 on node 104A stores data to storage device 108 the same as writing data to VD1 on node 104B or 104C. It is further noted that each storage device may be physically connected to more than one node. In this case, each node physically connected to the device has a different device driver that interfaces to the storage device.

In the illustrated embodiment, volume 1 (V1) of volume manager 105A is coupled to VD1 and VD2. In one embodiment, volume manager 105A may mirror these devices. In alternative embodiments, volume manager 105A may include other volumes coupled to other virtual devices. For example, a second volume manager 105A may be coupled to VD2 and VD3.

In nodes 104B and 104C, the volume managers (105B and 105C) and virtual disk systems (106B and 106C) operated in substantially the same manner as volume manager 105A and virtual disk system 106A. In the illustrated embodiment, volume 2 (V2) of volume manager 105B is coupled to VD2 and VD3 of virtual disk system 106B. Virtual disk system 106B directly accesses storage device 110 and accesses storage device 112 via communication interface 102 and node 104C. Volume 3 (V3) of volume manager 105C is coupled to VD2 and VD3 of virtual disk system 106C. Virtual disk system 106C directly accesses storage device 112 and accesses storage device 110 via communication interface 102 and node 104B.

Turning now to Figure 2, a block diagram of an alternative cluster configuration according to one embodiment of the present invention is shown. Cluster 200 includes a data communication link 102, three nodes

104A-104C, and three storage devices 108, 110 and 112. Components similar to those in Figure 1 are given the same reference numerals for simplicity. In Figure 2, the client interfaces to virtual disk system 106 rather than volume manager 105. The virtual disk system interfaces to the volume manager, which interfaces to one or more storage devices. In this configuration, volume manager 105 is layered below virtual disk system 106. For simplicity, only the operation of node 104A is discussed below. Nodes 104B and 104C operate in substantially the same manner.

In node 104A, the client interfaces to virtual disk system 106A. From the client's perspective, virtual disk system 106A appears as three separate storage devices. In Figure 2, the three virtual devices are labeled as virtual volumes (VV1, VV2 and VV3) to reflect the layering of the volume manager below the virtual disk system. From the client's perspective, virtual volumes behave like a storage device. For example, the virtual volume may use Unix device driver semantics. The client may access any of the three volumes of the cluster from virtual disk system 106A. Volume manager 105A interfaces to the storage devices. In the illustrated embodiment, volume 1 (V1) of volume manager 105A is coupled to storage devices 108 and 110. In one embodiment, volume 1 may mirror data on storage devices 108 and 110. From the perspective of virtual disk system 106A, volume 1 of volume manager 105A behaves like a storage device. For example, the volume may behave like a Unix device driver.

Virtual volume 2 (VV2) of virtual disk system 106B interfaces directly to volume 2 (V2) of volume manager 105B. Virtual volumes 1 and 3 communicate with volume 1 of node 104A and volume 3 of node 105C via data communication link 102. In a similar manner, virtual volume 3 of virtual disk system 106C interfaces directly to volume 3 of volume manager 105C. Virtual volumes 1 and 2 communicate with volume 1 of node 104A and volume 2 of node 105B via data communication link 102. In the illustrated embodiment, volume 2 of volume manager 105B and volume 3 of volume manager 105C are both physically connected to storage devices 110 and 112.

The volume manager may be layered either above or below the virtual disk system because both the volume manager and the virtual disk system behave like storage devices. Accordingly, it is transparent to the client whether it interfaces to the volume manager or the virtual disk system. In both embodiments, the client appears to have direct access to three reliable storage devices. Both the volume manager and the virtual disk system may interface directly to a storage device. Some volume managers may operate better when layered above the virtual disk device. For example, a cluster volume manager, such as the Veritas CVM, operates best when layered above the virtual disk system, while non-distributed volume managers, such as Solstice Disk Suite (SDS), may be required to operate below the virtual disk system. It is noted that a volume manager must be distributed to operate below the virtual disk system. It is further noted that a distributed volume manager, such as CVM, can manage the volumes (V1, V2 and V3) as though they are one volume, much like the virtual disk system manages the virtual disks on the nodes as though they are one device.

Turning now to Figure 3, a block diagram of a virtual disk system operating on two nodes of a cluster according to one embodiment of the present invention is shown. In the illustrated embodiment, each node includes a user portion and a kernel. The user portion of node 104A includes a cluster membership monitor (CMM) 310A, a cluster configuration database (CCD) 311A, a client 312A, a netdisk daemon (NDD) 314A, and a cluster transport interface daemon (CTID) 316A. The kernel of node 104A includes a netdisk driver (ND) 318A, a netdisk master

(NM) 320A, a cluster transport interface (CTI) 322A, a cluster connectivity monitor (CCM) 324A, a disk driver 326A and a network transport 328A. The user portion of node 104B includes a cluster membership monitor (CMM) 310B, a cluster configuration database (CCD) 311B, a netdisk daemon (NDD) 314B, and a cluster transport interface daemon (CTID) 316B. The kernel of node 104B includes a netdisk driver (ND) 318B, a netdisk master (NM) 320B, a cluster transport interface (CTI) 322B, a cluster connectivity monitor (CCM) 324B, a netdisk driver 326B and a network transport 328B.

In the illustrated embodiment, a volume manager is not included. As discussed above in reference to Figures 1 and 2, a volume manager may be implemented either above or below the virtual disk system. If the volume manager is implemented above the virtual disk system, client 312A interfaces to the volume manager, which in turn interfaces to ND 318A. Alternatively, if the volume manager is implemented below the virtual disk system, NM 320A interfaces to the volume manager, which in turn interfaces to disk driver 326A.

A configuration module called CTID 316A is a daemon that initializes a connection module called CTI 322A. When the configuration of the cluster changes or node 316A is initialized. CTID 316A queries CCD 311A to obtain configuration information. In one embodiment, configuration information indicates the number of links between the nodes of the cluster and the protocol associated with the links. In one embodiment, CTID 316A additionally queries CMM 310A to obtain membership information, such as a list of active nodes in the cluster. CTID 316A establishes connections over the links between the nodes and conveys the membership information and link information to CTI 322A. CTID 316A may communicate to CTI 322A via a private interconnect and may use an I/O control request.

The links identified by CCD 311A may be physical links or virtual links. For example, CCM 324A may manage a pair of physical links as one virtual link accessible by CTI 322A. CCM 324 is discussed in more detail below in reference to Figure 9.

CCD 311A is one instance of a distributed highly available cluster database. CCD 311 stores consistent data even in the presence of a failure. By storing mapping data in CCD 311, each node obtains the same mapping information even in the presence of a failure. CCD 311 is discussed in more detail in a co-pending, commonly assigned patent application entitled "Highly available Distributed Cluster Configuration Database" to Slaughter, *et al.*, filed on October 21, 1997, Serial No. 08/954,796.

CMM 310 is a distributed program that monitors the cluster membership. When the membership changes, CMM 310 detects that change and conveys new membership information to other resources in the cluster such as CTID 316A and NDD 314A. Examples of membership changes include a node joining or leaving the cluster. In one embodiment, CMM 310 outputs a configuration number unique to each configuration.

NDD 314A is a daemon that initializes ND 318A when a new device is opened or during reconfiguration. Reconfiguration may occur when a node joins or leaves the cluster, or when a node fails. In one embodiment, each virtual disk device is initialized separately. In one particular embodiment, a virtual disk device is initialized by a cluster when the device is opened by that cluster, or after a reconfiguration if the virtual disk device was open prior to the reconfiguration. In this manner, not all virtual disk devices are initialized after each reconfiguration.

In one embodiment, ND 318A stores a list of devices to be opened and a list of opened devices. When a client requests a device to be opened, ND 318A adds the device to the list of devices to be opened. NDD 314A queries the list of devices to be opened. If the list includes a device to open, NDD 314A queries CCD 311A to

obtain the mapping information for the identified device. NDD 314A may also query CMM 310A to obtain membership information, such as a list active nodes. NDD 314A conveys the mapping information and membership information to ND 318A. NDD 314A may communicate to ND 318A via a private interconnect and may use an I/O control request.

5 In one embodiment, the mapping information for a device identifies a primary and secondary node physically connected to a storage device and a disk device on those nodes corresponding to the storage device. Each pair of nodes and disks may be referred to as node/disk pairs. Based on the primary and secondary node/disk pair and the membership information, ND 318A may select a node to route a data access request for a device. Once ND 314A and CTI 322A have been initialize, the virtual disk system is ready to accept data access requests from
10 client 312A.

Client 312A accesses the virtual devices of the virtual disk system in the same manner as it accesses storage devices. From the client's perspective, it appears that each storage device, or volume, is physically connected to the node. In the illustrated embodiment, when client 312A accesses data from a storage device, it sends a data access request to ND 318A. In one embodiment, client 312A specifies the destination storage device,
15 the type of operation and the location to retrieve or store the data to ND 312A. The rest of the operation is transparent to client 312A. ND 318A, based on the mapping and current membership information, determines to which node to convey the data access request. In one embodiment, the mapping information obtained from CCD 311A includes a primary and secondary node physically connected to the storage device. ND 318A may route the data access request to the primary node if the primary node is active. Alternatively, if the primary node is not
20 active, then ND 318A may route the data access request to the secondary node. Which node is used to access the storage device is transparent to client 312A.

ND 318A conveys the data access request to CTI 322A and specifies to which node to convey the data access request. How CTI 322A transfers the data access request to the destination node is transparent to ND 318A and client 312A. In one embodiment, if the storage device is directly coupled to node 104A, ND 318A conveys the
25 data access request to NM 320A rather than CTI 322A. NM 320A conveys the data access request to disk driver 326A, which in turns accesses the storage device. In one embodiment, NM 320A is a portion of ND 318A that interfaces to disk driver 326A. Disk driver 326A interfaces to one or more storage devices physically connected to a node 104A.

CTI 322A manages a plurality of links. CTI 322A is one instance of the distributed program CTI 322.
30 CTI 322A may manage one or more links to the destination node of a data access request. For example, if the destination node for the data access request is node 104B, CTI 322A may manage three links to that node. CTI 322A may transport all the data to node 104B via one link or may distribute the data over the three links. CTI 322A may append a field to the data access request to identify the destination client at destination node. CTI 322B of node 104B may service multiple clients. The field appended to the message by CTI 322A identifies to which
35 client CTI 322B should route that data. For example, CTI 322A may append data to a data request received by ND 318A that specifies the destination client as ND 318B.

In one embodiment, CCM 324A manages two or more redundant physical links. From the perspective of CTI 322A, the redundant physical links appear as one logical link. CCM 324A exchanges messages over the physical links with CCM 324B. The two instances of CCM 324 reach agreement regarding which of the redundant

links are operational. CMM 324 may pick one operational physical link to transfer data. If that link fails, CCM 324 may detect the failure and transfer data on the alternate link. From the perspective of CTI 322, each logical link appears as one highly available link. In one embodiment, CCM 324A manages links to each node in the cluster. For example, CMM 324A may manage links to nodes 104B and 104C.

5 Network transport 328A performs the protocol functions over the links of data communication link 112. In one embodiment, a TCP/IP protocol is used over data communication link 112. In other embodiments, other protocols may be implemented. For example, a faster protocol such as Low Latency Connectivity Layer (LLCL), Message Passing Interface (MPI), or Low Overhead Communication (LOCO) may be used.

10 In node 104B, network transport 328B receives the data access request and transports the data using the appropriate protocol to CTI 322B. CTI 322B may partially decode the data access request to determine its destination client. In the illustrated embodiment, the data is routed to ND 318B. ND 318B may partially decode the data access request to determine the destination storage device. If the storage device is physically coupled to node 104B, ND 318B conveys the request to NM320B, which conveys the request to disk driver 326B. Disk driver 326B accesses the storage device. If the data access request is a read transaction, the requested data is routed back
15 to client 312A via the ND 318, CTI 322 and data communication link 112.

One feature of the virtual disk system according to one embodiment of the present invention is high availability. The virtual disk system is designed such that data access requests are reliably performed in the presence of a failure, such as a node failure. Towards this end, ND 318A stores a list of pending data access requests. If a data access request is not successfully completed, the virtual disk system retries the data access
20 request possibly to another node. The requesting node may detect an incomplete data access request by receiving a negative acknowledge signal or it may receive reconfiguration data indicating that a destination node is not active. When the data access request is successfully complete, it is removed from the list of pending data access requests.

For example, node 104B may be a primary node for a storage device and node 104C may be a secondary node for that storage device. When ND 318A conveys a data access request to the storage device, it may convey
25 the data access request to the primary node, which is node 104B. If node 104B is unable to successfully complete the data access request, for example if the storage device path between disk driver 326B and the storage device is non-functional, node 104A may receive a negative acknowledgement signal indicating that the data access request was not successfully completed. Node 104A may then resend the data access request to the secondary node, which is node 104C. Node 104A may store information indicating that node 104B is not able to communicate with the
30 storage device and subsequently send new data access requests to other nodes.

In an alternative example, node 104B may be non-operational. In one embodiment, the cluster membership data acquired by node 104A from CMM 310A may indicate that the node is not operational. Accordingly, ND 318A may route data access requests to the secondary node. In the above manner, data access requests are successfully completed even in the presence of a failure.

35 Turning now to Figure 4, a block diagram illustrating the initialization of a netdisk driver is shown according to one embodiment of the present invention. Figure 4 illustrates the initialization of ND 318A in node 104A. The initialization of other netdisk drivers in the cluster may be performed in a substantially similar manner.

In one embodiment, prior to accessing a storage device, the storage device is opened. For example, an open command may be executed that causes the storage device to be initialized. Similarly each virtual device on

each node may be opened prior to accessing it. Client 312A outputs a command to ND 318A to open a virtual device. ND 318A stores the device to be opened in a list. In one embodiment, NDD 314A periodically queries the list to determine which devices to initialize. In an alternative embodiment, ND 318A may output a signal to NDD 314A indicating that a device needs to be initialized. NDD 314A queries CCD 311A to obtain mapping information for the device to be opened, and queries CMM 310A for current membership information. NDD 314A conveys the mapping and membership information to ND 318A. ND 318A stores the mapping and membership information to a configuration file. ND 318A uses the mapping and membership data stored in the configuration file to determine the routing of data access requests to nodes. ND 318A then notifies client 312A that the device has been opened.

In one embodiment, the mapping information for each virtual device includes: the name of the virtual device, a primary node, the name of the storage device at the primary node (i.e., the name of the device that corresponds to the storage device), a secondary node and the name of the storage device at the secondary node. The mapping information may additionally include an identification number for the virtual device and a cluster-unique name for the storage device.

ND 318A additionally stores a reconfiguration number associated with the mapping and membership data. The reconfiguration number is obtained from CCM 310A. ND 318A uses the reconfiguration number to determine whether its current membership data is up to date with respect to the most recent configuration.

In one embodiment, when the configuration of the cluster changes, CMM 310A notifies NDD 314A of the new membership information. For example, if a node failure is detected, CMM 310A will notify NDD 314A that a reconfiguration has occurred and convey the new membership data to NDD 314A. NDD 314A conveys the new membership information to ND 318A, which uses the new membership information in conjunction with the mapping information to route future data access requests.

In one embodiment, a filesystem manages the virtual disks on a node. This filesystem may be called a netdisk filesystem (NDFS). NDFS is configured to create a special device file for virtual disks when a node opens the virtual disk. The special device file represents the virtual disk in the operating system.

In operating systems, such as the UNIX operating system, devices may be treated as files. The file associated with a device (called a device file or a special device file) is normally created by an initialization program that runs during the boot-up phase of the operating system. The initialization program determines the physical devices attached to the computer system and creates device files corresponding to those physical devices. In one embodiment, virtual devices are initialized the first time they are accessed rather than during boot-up. This situation and the fact that the virtual disk may not be physically connected to the node means that the device files for the virtual disks may not be created during initialization. Because the virtual disks preferably are accessible like other devices, NDFS is configured to create device files for the virtual devices when they are first opened. In one embodiment, a device file is only created the first time a node opens a virtual device. Subsequent opens of the virtual device do not cause device files to be created.

In one embodiment, NDFS detects a command to open a virtual device. If this is the first time the virtual device has been opened, NDFS sends a creation request to ND 318A. In one embodiment, NDFS has a private interface to ND 318A. ND 318A stores the virtual device to create in a list. The list may be the same list used to store devices to open or may be a separate list for devices to create. NDD 314A may periodically query the list to

determine which devices to create or ND 318A may output a signal to NDD 314A indicating a device needs to be created. NDD 314A queries CCD 311A to obtain permission data for the device to be opened. NDD 314A conveys the permission data to ND 318A which in turn conveys the permission data to NDFS. NDFS will create the device file for the device with the permission data received from CCD 311A. In one embodiment, the device is opened after the device file is created using a normal device open procedure as discussed above. Subsequent opens of the same device by the same node may result in a normal open operation without the need for NDFS to be involved. Accordingly, a performance penalty is only incurred the first time a device is opened. Subsequent commands to open the device are performed in the same manner as the opening of any other device.

Turning now to Figure 5, a block diagram illustrating the initialization of a cluster transport interface according to one embodiment of the present invention is shown. Figure 5 illustrates the initialization of CTI 316A in node 104A. The initialization of other cluster transport interfaces in the cluster may be performed in a substantially similar manner.

In one embodiment, prior to transferring data over data communication link 102, CTID 316A establishes connections over the available links. During initialization, CTID 316A queries CMM 310A for data identifying the current cluster membership and queries CCD 311A for data identifying which links are connected to which nodes. In one embodiment, CCD 311A stores additional information about the links such as the transfer protocol of the links. CTID 316A establishes connections over the available links and passes the link information and membership data to CTI 322A. In one embodiment, CTID 316A establishes TCP/IP connections over the available links.

CTI 322A interfaces to network transport 328A to exchange data to other instances of CTI 322. In one embodiment, network transport 328A interfaces to CCM 324A, which manages one or more redundant links. When CTI 322A receives a data access request destined for a particular node, it determines which connections connect the requesting node to the destination node. CTI 322A determines on which connection(s), to transport the data to the destination node. For example, if CTI 322A manages connections over three links to node 104B and it receives a data access request destined for that node, it may transfer all the data via one connection or it may transfer a portion of the data over each of the three connections.

When the cluster is reconfigured, CMM 310A notifies CTID 316A of the event. CTID 316A obtains the new membership data from CCD 311A and conveys the new membership data and a new configuration number to CTI 322A. Additionally, CTID 316A may obtain link data from CCD 311A and conveys that data to CTI 322A. CTID 322A may modify the connections when a reconfiguration occurs. For example, CTID 322A may establish connections over links to new nodes in the cluster, or may abandon connections to nodes that leave the cluster.

Turning now to Figure 6, a flowchart diagram illustrating the operation of a virtual disk system according to one embodiment of the present invention is shown. In step 612, a netdisk driver is initialized. The initialization of the netdisk driver is discussed in more detail in reference to Figure 7. In step 614, a cluster transport interface is initialized. The initialization of the cluster transport interface is discussed in more detailed in reference to Figure 8. In step 616, the netdisk driver receives a data access request from a client. In step 617, the netdisk driver stores the data access request and any other data necessary to re-issue the data access request if it is not successfully completed.

In step 618, the netdisk driver that receives the data access request determines whether the destination device is physically connected to the requesting node. If the destination device is physically connected to the

requesting node, then in step 620 the netdisk driver performs the data access request on the storage device. Alternatively, if the storage device is not physically connected to the requesting node, then in step 620 the netdisk driver detects a node to which to convey the data access request. In one embodiment, the netdisk driver stores mapping information identifying a primary and secondary node for each storage device. In one particular
5 embodiment, the netdisk driver selects the primary or secondary node based upon membership data and/or previous unsuccessful data access requests. In step 624, the netdisk driver conveys the data access request to the selected destination node via the cluster transport interface.

In step 626, the cluster transport interface selects one or more connections to transfer data to the destination node by the netdisk driver. In step 628, the cluster transport interface conveys the data access request to
10 the destination node via the selected connection(s). In step 630, the cluster transport interface at the destination node receives the data access request and determines the destination client, which in the instant example is the netdisk driver, or more particularly the netdisk master. In step 632, the netdisk master receives the data access request and accesses the destination storage device. In step 634, the cluster transport interface of the destination node returns an acknowledge or not acknowledge signal to the requesting node. If the data access request is a read
15 request, the requested data may also be returned to the requesting node.

In parallel with the transfer of the data access request, in step 638, the requesting node waits for a status signal from the destination node. The status signal may take the form of an acknowledge or a not acknowledge signal. In step 640, it is determined whether or not an acknowledge was received. If an acknowledge signal is received, then operation continues at step 616. Alternatively, if a not acknowledge signal is received, then in step
20 642 an alternate node to convey the data access request is selected and operation continues at step 624.

Turning now to Figure 7, a flowchart diagram illustrating the initialization of a netdisk driver according to one embodiment of the present invention is shown. In step 712, the netdisk daemon queries that netdisk driver for devices to open. In decisional step 714, it is determined whether any devices need to be opened. If no devices need to be opened, execution continues at step 712. Alternatively, if the netdisk daemon detects a device to open, then in
25 step 716 the netdisk daemon queries the cluster configuration database for mapping data. The mapping data may identify node/disk pairs mapped to a virtual device. In step 718, the netdisk daemon queries the cluster membership monitor for membership data.

In step 720, the netdisk daemon conveys the mapping and membership data to the netdisk driver. In step 722, the netdisk driver updates the mapping information for the device, updates that membership information and
30 records the reconfiguration number. In step 724, the netdisk driver notifies the client that the requested device is open.

Turning now to Figure 8, a flowchart diagram illustrating the initialization of a cluster transport interface according to one embodiment of the present invention is shown. In step 812, a cluster transport interface daemon receives an indication of a configuration change. Alternatively, the cluster transport daemon may receive an
35 indication of a system initialization. In step 814, the cluster transport interface daemon queries the cluster configuration database for link information. In one embodiment, link information may include the number of links between nodes within a cluster, which links are coupled to which nodes, and information such as the protocol used by the links. In step 816, the cluster transport interface daemon queries the cluster membership monitor for membership information.

In step 818, the cluster transport interface establishes connections over the links. In step 820, the cluster transport interface daemon conveys the link and membership information to the cluster transport interface. The cluster transport interface is then ready to accept data access requests or other messages.

Turning now to Figure 9, a block diagram of the cluster transport interface according one embodiment of present invention is shown. A cluster transport interface is one example of a data transport system. Figure 9 includes three instances of a cluster transport interface (322A-322C), three TCP/IP interfaces (912A-912C), and eight cluster connection monitors (914A-914H). CTI 322 is a distributed software program that provides a facility for passing messages between nodes. The messages may include control messages and data blocks.

The instances of cluster transport interface 322 pass data between client programs. For example, CTI 322A may receive a message from a netdisk driver that is a client to CTI 322A. In one embodiment, the message specifies its destination node and a disk device on that node. CTI 322A determines which links are connected to the destination node and conveys the message over one of those links. The cluster transport interface at the destination node receives the data access request, determines the destination client and conveys the data to the destination client. For example, CTI 322A may route a data access request from the netdisk driver in node 104A to the netdisk driver in node 104B. CTI 322B receives the data access request, determines the destination client and conveys the data access request to the netdisk driver in node 104B. From the perspective of a client, CTI 322A appears as one virtual link to the destination node.

In the illustrated embodiment, CTI 322 uses TCP/IP for transferring data to other nodes. CTID 316A automatically establishes a TCP/IP connection over each link during initialization. CTI 322 conveys a message to TCP/IP 912A which transfers the message to the appropriate instance of CCM 914. CTI 322A, however, is not dependent upon any particular data transfer protocol. By modifying TCP/IP 912 and/or CCM 914, CTI 322 may interface to any data transport interface or transfer protocol.

In one embodiment, CTI 322A allocates memory for storing messages and data received from other nodes and deallocates the memory when the data are no longer required by a client. In one embodiment, CTI 322 uses a call-back function to indicate to a client that data have been received. For example, CTI 322A may convey a read request to node 104B. When CTI 322A receives the requested data it uses a call-back function to the requesting client to indicate the data are available.

Cluster connection monitor (CCM) 914 manages two or more physical links as one logical link. In the illustrated embodiment, a pair of instances of CCM 914 manages two links. In alternative embodiments, a pair of instances of CCM 914 may manage more links. Pairs of physical links connect one node in the cluster to another node. For example, links 916A couple node 104A to node 104B, and links 916B couple node 104A to node 104C. In one embodiment, the links are handled as redundant links by CMM 914. Data is transferred on one link until a failure of that link is detected and then data is transferred on the other link.

CCM 914 determines which links are operational and detects failures by exchanging messages, sometimes called heartbeat messages, over both physical links. For example, CCM 914A and CCM 914E exchange heartbeat messages to determine whether physical links 916A are operational. The two instances of CCM 914 select one of the physical links as the primary link. If the primary link fails, CCM 916 detects the failure and begins transferring data on the other physical link. In one particular embodiment, CCM 916 exchanges Unreliable Data Protocol (UDP) messages across a physical link to determine whether the link is operational.

From the perspective of CTI 322, each pair of physical links managed by CCM 914 appears as one logical link. Accordingly, the data transferred by CTI 322A may be transferred on one of the two physical links transparent to CTI 322A.

In the illustrated embodiment, three logical links (916B-916D) connect node 104A to node 104C. CTI 322A determines on which of the three links to transfer the data. In one embodiment, CTI 322A may transfer all the data on one logical link. In alternative embodiment, CTI 322A may transfer a portion of the data on each logical link. As noted above, it is transparent to the client on which or how many logical links the data are transferred.

Turning now to Figure 10, a diagram illustrating device permissions according to one embodiment of the present invention is shown. The permission data are shown in the context of a listing of a directory. A similar listing may be obtained by performing an "ls -l" command on a directory that lists raw virtual disk devices. It is noted that the device permissions are related to the devices themselves, not to the files or directories on those devices. The raw devices (i.e., devices with no filesystem or files on them) are treated as files for permission purposes.

Field 1012 includes ten characters. The first character is either a "d", which identifies a directory, or a "-", which identifies a device. The next nine characters are three groups of three characters. Each group represents the permission modes for a owner, a group and others, respectively. The permission modes include read (r), write (w) and execute (x). One character in each group represents each permission mode. If a letter representing the permission mode is present, then the associated user has that permission. Alternatively, if a "-" is present, the associated user does not have that permission. For example, if a user has the following permissions "rwx" then the user has read, write and execute permission. Alternatively, if the user has the following permissions "r--" then the user has read permission, but not write or execute permission. The first group of three characters represents the permissions for the owner of the device. The second group of three characters represents the permissions for a group associated with the device. The last group of three characters represents the permissions for other users. Owners and groups are discussed in more detail below. For example, if the permissions in field 1012 are "drwx--x-x", the field represents a directory, the owner has read, write and execute permission, and the group and others have execute permission only.

Field 1016 identifies the owner of the entry. The owner is the user that created the device. Field 1018 identifies a group of related users. Groups are defined within the operating system. Field 1018 associates one of the defined groups with the device. Other users that are neither the owner nor within the selected group. As discussed above, different permissions may be defined for the owner, group and other users.

Field 1022 identifies the date and time of the last modification of the device. If the last modification is within the current calendar year, the month, day and time are specified. Alternatively, if the last modification is not within the current calendar year, the month, day and year are specified. Field 1024 identifies the name of the device.

To ensure consistent permission data among the nodes of the cluster, the permission data may be stored in a highly available database. In one embodiment, multiple nodes within a cluster have representations of a device. To maintain consistent permission data among the nodes even in the presence of a failure, the permission data is stored in a cluster configuration database (CCD).

In one embodiment, when a node first opens a virtual device, the permission data for that device are read from the CCD and a device file is created with the permission data. In one embodiment, the device file is only created the first time a virtual device is opened by a node. In one embodiment, a filesystem operating on each node includes a daemon that queries the CCD for permission data of the device. The daemon returns the permission data to the filesystem, which creates a special device file with those permissions. Because the CCD may be queried by any node of the cluster and returns consistent information even in the presence of a failure, all nodes will have consistent permission data.

Turning now to Figure 11, a flowchart diagram illustrating the storage and access of consistent permission data according to one embodiment of present invention is shown. In step 1112, permission data are stored to a highly available database. In one embodiment, the permission data include device permissions, the owner of the device, and the group associated with the device. In step 1114, a first node opens a device on a first node and accesses the permission data from the highly available database. In step 1115, the node opens a special device file for the device including the permission data. In step 1116, a second node opens a device corresponding to the same physical device on a second node and accesses the permission data. In step 1117, the node opens a special device file for the device including the permission data on the second node. Because the highly available database returns consistent data, the nodes receive consistent permission data.

Turning now to Figure 12, a flowchart diagram illustrating the update of a configuration mapping according to one embodiment of the present invention is shown. In step 1212, an indication that an update is pending is provided to the nodes. In step 1214, the nodes suspend data access requests to the storage devices. In step 1216, the nodes wait for outstanding data access requests to complete. In step 1218, the nodes invalidate an internal representation of a mapping of virtual disks to storage devices. In step 1220, the nodes output acknowledge signals indicating that the internal mapping representations have been invalidated, data access requests have been suspended, and outstanding data access requests have completed. In step 1222, the system waits for acknowledge signals from all active nodes. In step 1224, the system updates its mapping. In step 1226, the system outputs an indication that the update is complete. In step 1228, the nodes request an updated version of the mapping. In step 1230, the nodes resume sending data access requests to storage devices.

In one embodiment, the update procedure is coordinated by a cluster configuration database (CCD). To prevent errors, the mapping should be updated consistently among all the nodes. The CCD notifies the nodes of a pending update and notifies the nodes that the update is complete via a synchronization command. In one embodiment, the synchronization command is run whenever a row in the CCD is modified. The command to run during modification of a row in the CCD may be specified in a format row associated with the data stored in the CCD. The synchronization command may be run in parallel on all the nodes of the cluster. In one embodiment, a netdisk synchronization command is run when the netdisk mapping is modified. A different invocation of the netdisk synchronization command may be run depending upon the type of the modification. The CCD outputs a first synchronization command prior to modifying the mapping. A second synchronization command may be run after the database is updated.

In one embodiment, if an acknowledge signal is not received from all nodes, the cluster will suspend the update and output a cancel signal. In one embodiment, the cancel signal causes the node to revalidate the internal mapping representations and continue operating.

In the above described manner, the configuration of a cluster can be modified while the cluster is operating without losing data. The data access requests in the system may be delayed, but they will proceed without error. The above described reconfiguration procedure also allows connections to be reconfigured without losing data. For example, a storage device can be disconnected from one node and reconnected to another node. The physical
5 reconfiguration may occur between steps 1222 and 1224. Further, the reconfiguration is transparent to the client except for a delay. Another application of the above described reconfiguration is to change the mapping (or administration) of the volume manager during operation.

Numerous variations and modifications will become apparent to those skilled in the art once the above disclosure is fully appreciated. It is intended that the following claims be interpreted to embrace all such variations
10 and modifications.

WHAT IS CLAIMED IS:

1. A distributed computing system comprising:
 - 5 a first node, a second node, and a communication link coupled between said first node and said second node; and
 - a storage device coupled to said first node wherein said second node is configured to access said storage device using a virtual disk system that is configured to make said storage device appear to be
 - 10 coupled to said second node;
 - wherein said virtual disk system includes a driver operating on said second node and a master operating on said first node, wherein when said second node accesses a virtual disk of said virtual disk system, said driver is configured to send a data request to said master, said master is configured to access
 - 15 data from said storage device and said master is configured to convey a response to said driver via said communication link;
 - wherein said driver is configured to store a copy of said data request until said driver receives said response and said driver is configured to resend said data request if said driver fails to receive said response.
 - 20
2. The distributed computing system of Claim 1 further comprising a third node coupled to said data communication interface and said storage device wherein said driver is configured to resend said data request to a second master on said third node if said driver fails to receive said response, wherein said second master is
- 25 configured to access data for said storage device.
3. The distributed computing system of Claim 1 wherein if said driver fails to receive said data due to a failure of said first node or of a communication link, said driver is configured to resend said data request after said distributed computing system reconfigures.
- 30
4. The distributed computing system of Claim 3 wherein said resent data request is sent to a third node which is coupled to said storage device.
5. The distributed computing system of Claim 4 wherein said virtual disk system maintains a primary node
- 35 and an alternate node to communicate with said storage device, wherein said second node is configured to send said data request to said primary node and to resend said data request to said alternate node if said second node fails to receive a response.

6. The distributed computing system of Claim 5 wherein said first node is said primary node and said third node is said alternate node.
7. A distributed computing system comprising:
- 5 a first node, a second node, and a communication link coupled between said first node and said second node; and
- 10 a storage device coupled to said first node wherein said second node is configured to access said storage device using a virtual disk system that is configured to make said storage device appear to be coupled to said second node;
- wherein said virtual disk system is configured to include a mapping of a virtual disk to said storage device, wherein said first node and said second node are configured to receive consistent mapping data.
- 15 8. The distributed computing system of Claim 7 wherein said mapping identifies a node coupled to said storage device and a disk device on said node corresponding to said storage device, and said mapping is stored in a highly available database accessed by said first node and said second node.
- 20 9. The distributed computing system of Claim 8 wherein said highly available database is a cluster configuration database.
10. The distributed computing system of Claim 7 wherein said mapping includes data indicative of a primary node coupled to said storage device and an alternate node coupled to said storage device.
- 25 11. The distributed computing system of Claim 7 wherein data access requests are configured to be suspended when said mapping is updated.
12. The distributed computing system of Claim 11 wherein said mapping is configured to be updated when a node joins or leaves a cluster.
- 30 13. The distributed computing system of Claim 12 wherein nodes query a database for a new mapping when data access requests are resumed.
- 35 14. The distributed computing system of Claim 13 whereby the configuration of the distributed computing system may be updated during operation.

15. A distributed computing system comprising:

a first node, a second node, and a communication link coupled between said first node and said second node; and

a storage device coupled to said first node wherein said second node is configured to access said storage device using a virtual disk system that is configured to make said storage device appear to be coupled to said second node;

wherein permission data of said storage device are consistent between said first node and said second node.

16. The distributed computing system of Claim 15 wherein said permission data are configured to be stored in a highly available database accessed by said first node and said second node.

17. The distributed computing system of Claim 16 wherein said highly available database is a cluster configuration database.

18. A distributed computing system comprising:

a first node, a second node, and a communication link coupled between said first node and said second node;

a first storage device coupled to said first node wherein said second node is configured to access said first storage device using a virtual disk system that is configured to make said first storage device appear to be coupled to said second node;

a second storage device coupled to said second node wherein said first node is configured to access said second storage device using a virtual disk system that is configured to make said second storage device appear to be coupled to said first node; and

a volume manager layered above said virtual disk system or below said virtual disk system.

19. The distributed computing system of Claim 18 wherein said volume manager is layered below said virtual disk system, wherein said virtual disk system is configured to access a volume, and said volume is configured to access said first storage device or said second storage device.

20. The distributed computing system of Claim 19 wherein said volume appears as a storage device to said virtual disk system.

21. The distributed computing system of Claim 20 wherein said volume is a highly available storage device.
22. The distributed computing system of Claim 18 wherein said volume manager is layered above said virtual disk system, wherein a client may access a volume which is configured to access a virtual disk of said virtual disk system.
23. The distributed computing system of Claim 22 wherein said volume is configured to access two or more virtual disks.
24. A distributed computing system comprising:
- a first node, a second node, and a communication link coupled between said first node and said second node; and
- a storage device coupled to said first node wherein said second node is configured to access said storage device using a virtual disk system that is configured to make said storage device appear to be coupled to said second node;
- wherein said virtual disk system includes a driver operating on said second node and a master operating on said first node, wherein when said second node accesses a virtual disk corresponding to said storage device, said driver is configured to send a data request to said master, said master is configured to access data from said storage device and said master is configured to convey said data to said driver via said communication interface,
- wherein said driver is configured to store a copy of said data request until said driver receives said data and said driver is configured to resend said data request if said driver fails to receive said data, said virtual disk system is configured to include a mapping of a virtual disk to said storage device, wherein said first node and said second node are configured to receive consistent mapping data in the event of a node failure; and wherein permission data of said storage device are consistent between said first node and said second node.
25. The distributed computing system of Claim 24 further comprising a third node coupled to said data communication interface and said storage device wherein said driver is configured to resend said data request to said third node if said driver fails to receive said data.
26. The distributed computing system of Claim 25 wherein said mapping includes data indicative of a primary node coupled to said storage device and a secondary node coupled to said storage device.

27. The distributed computing system of Claim 26 wherein said mapping and said permission data are configured to be stored in a highly available database accessible by said first node and said second node.
28. A distributed computing system comprising:
- 5 a first node, a second node, a storage device and a communication link coupled to said first node and said second node, wherein said first node and said second node access said storage device;
- 10 wherein said storage device has permission data associated with said storage device and said permission data are stored in a highly available distributed database accessible by said first node and said second node; when a particular node opens said storage device, said particular node creates a device file with said permission data for said storage device, whereby said first node and said second node obtain consistent permission data in the presence of a node failure.
- 15 29. The distributed computing system of Claim 28 wherein said permission data includes an owner, a group and permission modes for said owner and said group.
30. The distributed computing system of Claim 29 wherein said permission modes include read, write and execute.
- 20 31. The distributed computing system of Claim 28 wherein said highly available database is a cluster configuration database.
32. The distributed computing system of Claim 28 wherein said storage device is a disk device.
- 25 33. The distributed computing system of Claim 28 wherein said first node directly accesses said storage device and said second node accesses said storage device via said communication link.
34. The distributed computing system of Claim 28, wherein said device file is created a first time said storage device is opened by a particular node.
- 30 35. A method of maintaining consistent permission data among a plurality of nodes in a distributed computing system that includes a storage device, said method comprising:
- 35 storing said permission data to a highly available distributed database;
- a first of said plurality of nodes opening a device and accessing said highly available database to obtain permission data for said device;

a second of said plurality of nodes opening said device and accessing said highly- available database to obtain said permission data for said device;

whereby said first node and said second node obtain consistent permission data in the presence of a failure.

5

36. The method of Claim 35 wherein said nodes opening said device create a device file for said storage device, wherein said device file includes said permission data.

10

37. The method of Claim 36 wherein said permission data includes an owner, a group and permission modes for said owner and said group.

38. The method of Claim 37 wherein said permission modes include read, write and execute.

39. The method of Claim 35 wherein said highly available database is a cluster configuration database.

15

40. The method of Claim 35 wherein said storage device is a disk device.

41. The method of Claim 35 wherein said first node directly accesses said storage device and said second node accesses said storage device via said data communication link.

20

42. The method of Claim 36 wherein said device file is created a first time said storage device is opened by a particular node.

25

43. A computer-readable storage medium comprising program instructions for maintaining consistent permission data among a plurality of nodes in a distributed computing system, wherein said program instructions execute on said plurality of nodes of said distributed computing system and said program instructions are operable to implement the steps of:

storing said permission data to a highly available coherent database;

30

a first of said plurality of nodes opening a device and accessing said highly available database to obtain permission data for said device;

a second of said plurality of nodes opening said device and accessing said highly- available database to obtain said permission data for said device;

35

whereby said first node and said second node obtain consistent permission data in the presence of a failure.

44. The medium of Claim 43 wherein said file for said storage device, wherein said device file includes said permission data.

40

45. The medium of Claim 44 wherein said permission data includes an owner, a group and permission modes for said owner and said group.

46. The medium of Claim 45 wherein said permission modes include read, write and execute.

47. The medium of Claim 43 wherein said highly available database is a cluster configuration database.

48. The medium of Claim 44 wherein said device file is created a first time said storage device is opened by a particular node.

49. A distributed computing system comprising:

one or more nodes coupled to a communication link and one or more storage devices coupled to said one or more nodes wherein said one or more nodes include one or more devices; and

a highly available database accessible by said one or more nodes, wherein said database provides coherent data to said one or more nodes in the presence of a failure;

wherein a mapping of said one or more devices to said one or more storage devices is stored in said highly available database;

wherein when said mapping is updated, said one or more nodes stop issuing data requests to said one or more storage devices prior to said highly available database updating said mapping, and said one or more nodes resume issuing data requests when said mapping is updated.

50. The distributed computing system of Claim 49 wherein said nodes send an acknowledge signal to said highly available database when said nodes have stopped issuing data requests.

51. The distributed computing system of Claim 50 wherein said one or more nodes wait for outstanding data requests to complete prior to sending said acknowledge signal.

52. The distributed computing system of Claim 51 wherein said highly available database updates said mapping after receiving said acknowledge signals.

53. The distributed computing system of Claim 52 wherein said highly available database outputs a first synchronization signal to indicate a pending update of said mapping.

54. The distributed computing system of Claim 53 wherein said highly available database outputs a second synchronization signal to indicate said mapping is updated.

55. The distributed computing system of Claim 54 wherein said first synchronization commands and said second synchronization commands are issued to said one or more nodes concurrently.

56. The distributed computing system of Claim 55 wherein said highly available database waits for an
5 acknowledge signal from each active node prior to updating said mapping.

57. The distributed computing system of Claim 56 whereby said computer system reconfigures said connections between said one or more nodes and said one or more storage devices without losing or corrupting data.

10

58. A method of updating a mapping of nodes to storage devices, comprising:

storing said mapping in a highly available database wherein said database is accessible by said nodes and provides coherent data to said nodes in the presence of a failure;

15

said database outputting an indication to said nodes that an updated mapping is pending;

said nodes suspending data requests to said storage devices;

20

said nodes waiting for outstanding data requests to complete;

said nodes invalidating an internal representation of said mapping;

said nodes outputting acknowledge signals to said database;

25

said database waiting for said acknowledge signals from active nodes;

said database updating said mapping;

said database outputting an indication to said nodes that said update is complete;

30

said nodes requesting an updated version of said mapping from said database; and

said nodes resuming sending said data requests to said storage devices.

59. The method of Claim 58 wherein if said database does not receive an acknowledge signal from an active node, said database outputs a cancel indication to said nodes to reactivate said internal representation of said mapping.

60. The method of Claim 58 wherein said indication to said nodes are synchronization signals from said database.

61. The method of Claim 60 wherein said synchronization signals are issued to said one or more nodes concurrently.

62. The method of Claim 61 whereby said mapping is updated without losing or corrupting data.

63. A method of updating a mapping of nodes to storage devices, comprising:

storing said mapping in a highly available database wherein said database is accessible by said nodes and provides coherent data to said nodes in the presence of a failure;

said database outputting an indication to said nodes that an updated mapping is pending;

said nodes suspending data requests to said storage devices;

said database updating said mapping;

said database outputting an indication to said nodes that said update is complete; and

said nodes resuming sending said data requests to said storage devices.

64. The method of claim 63 further comprising prior to said database updating said mapping:

said nodes outputting acknowledge signals to said database; and

said database waiting for said acknowledge signals from said active nodes.

65. The method of claim 64 further comprising prior to said nodes outputting acknowledge signals to said database:

said nodes waiting for outstanding data requests to complete; and

said nodes invalidating an internal representation of said mapping.

66. The method of Claim 65 further comprising prior to said nodes resuming sending said data requests to said storage devices said nodes requesting an updated version of said mapping from said database.

67. The method of Claim 63 wherein if said database does not receive an acknowledge signal from an active node, said database outputs a cancel indication to said nodes to reactivate said internal representation of said mapping.

5 68. The method of Claim 67 wherein said indication to said nodes is a synchronization command from said database.

69. The method of Claim 68 wherein said synchronization signal is issued to said one or more nodes concurrently.

10

70. The method of Claim 69 whereby said mapping is updated without losing or corrupting data.

71. A computer-readable storage medium comprising program instructions for updating a mapping of nodes to storage devices, wherein said program instructions execute on a plurality of nodes of a distributed computing system and said program instructions are operable to implement the steps of:

15

storing said mapping in a highly available database wherein said database is accessible by said nodes and provides coherent data to said nodes in the presence of a failure;

20

said database outputting an indication to said nodes that an updated mapping is pending;

said nodes suspending data requests to said storage devices;

said nodes waiting for outstanding data requests to complete;

25

said nodes invalidating an internal representation of said mapping;

said nodes outputting acknowledge signals to said database;

30

said database waiting for said acknowledge signals from active nodes;

said database updating said mapping;

said database outputting an indication to said nodes that said update is complete;

35

said nodes requesting an updated version of said mapping from said database; and

said nodes resuming sending said data requests to said storage devices.

72. A data transport system of a distributed computing system including a first node, a second node, a third node and a data communication link comprising:

5 a configuration module configured to determine a number of active nodes of said distributed computing system and a number of links between said active nodes, and to establish connections over said links;

10 a connection module configured to receive data indicative of said number of active nodes and said number of links from said configuration module, to receive a request from a client to transfer data to a first active node, and to convey said data to said first active node via one or more of said links;

15 wherein when said number of active nodes changes, said configuration module notifies said connection module of said change and said connection module is configured to reestablish said connections to said active nodes transparent to said client.

73. The data transport system of Claim 72 wherein said configuration module receives data indicative of said number of said links from a highly available database.

20 74. The data transport system of Claim 73 wherein said highly available database is accessible by all of said nodes, whereby each node receives consistent data.

75. The data transport system of Claim 74 wherein said highly available database stores consistent data in the presence of a node failure.

25 76. The data transport system of Claim 72 wherein said configuration module is a daemon.

77. The data transport system of Claim 76 wherein said connection module is a kernel module.

30 78. The data transport system of Claim 72 wherein said configuration module and said connection module communicate via a private interface.

79. The data transport system of Claim 72 wherein said data communication link provides an interface between an instance of said connection module running on said first node and instance of said connection module running on said second node.

35 80. The data transport system of Claim 79 wherein said data communication link comprises multiple physical links between said first node and said second node, and said configuration module manages said multiple physical links as one virtual link.

81. The data transport system of Claim 72 wherein said data transport system services multiple clients.
82. The data transport system of Claim 81 wherein said multiple clients send and receive messages between said nodes.
- 5 83. The data transport system of Claim 82 wherein said configuration module notifies a client of a message received from another active node via a call-back function.
84. The data transport system of Claim 72 wherein data transferred by said data transport system comprises
10 messages.
85. The data transport system of Claim 84 wherein a message includes both a control message and a data portion.
- 15 86. The data transport system of Claim 72 wherein said connection module allocates and frees storage space for messages.
87. The data transport system of Claim 86 wherein a client notifies said connection module when data from a message is no longer required and said connection module frees storage space associated with said message.
- 20 88. A method of transporting data in a distributed computing system comprising a plurality of nodes and a data communication bus, the method comprising:
- determining physical resources in said distributed computing system, wherein said physical resources
25 include active nodes of said distributed computing system and active links between said active nodes;
- establishing a connection over said active links;
- 30 receiving a data access request to convey data to a first of said active nodes;
- conveying said data over one or more of said active links to said first active node;
- determining that said physical resources have changed; and
- 35 reestablishing connections to said changed physical resources;
- wherein said determination of changed resources and said reestablishing of links are transparent to a client.

89. The method of Claim 88 wherein multiple links between active nodes are handled as one logical link.
90. The method of Claim 88 wherein said determining of physical resources is performed by a daemon module.
- 5 91. The method of Claim 90 wherein establishing a connection over said active links is performed by a daemon module.
- 10 92. The method of Claim 91 wherein conveying said data to said active nodes is performed by a kernel module.
93. The method of Claim 88 wherein multiple clients are supported, wherein said data conveyed to said active nodes includes an identification of a client that requested the data access request.
- 15 94. The method of Claim 88 wherein said conveyed data includes a control portion and a data portion.
95. The method of Claim 88 further comprising:
- allocating memory space to store said data conveyed to an active node; and
- 20 freeing said memory space.
96. The method of Claim 89 further comprising notifying a client at a destination node of the receipt of data directed to said client.
- 25 97. The method for Claim 89 wherein determining physical resources includes accessing a highly available database that stores a list of physical resources.
98. The method of Claim 97 wherein said highly available database is accessible by said active nodes,
- 30 whereby said active nodes have consistent configuration data.
99. A computer-readable storage medium comprising program instructions for transporting data in a distributed computing system comprising a plurality of nodes and a data communication link, wherein said program instructions execute on a said plurality of nodes of said distributed computing system and said
- 35 program instructions are operable to implement the steps of:
- determining physical resources in said distributed computing system, wherein said physical resources include active nodes of said distributed computing system and active links between said active nodes;

establishing a connection over said active links;

receiving a data access request to convey data to a first of said active nodes;

5 conveying said data over one or more of said active links to said first active node;

determining that said physical resources have changed; and

reestablishing connections to said changed physical resources;

10

wherein said determination of changed resources and said reestablishing of connections are transparent to
a client.

100. The computer-readable storage medium of Claim 99 wherein multiple links between active nodes are
15 handled as one logical link.

101. The computer-readable storage medium of Claim 99 wherein said determining of physical resources is
performed by a daemon module.

20 102. The computer-readable storage medium of Claim 101 wherein establishing a connection over said active
links is performed by a daemon module.

103. The computer-readable storage medium of Claim 102 wherein conveying said data to said active nodes is
performed by a kernel module.

25

104. The computer-readable storage medium of Claim 99 further comprising:

allocating memory space to store said data conveyed to an active node; and

30 freeing said memory space.

105. The computer-readable storage medium of Claim 99 further comprising notifying a client at a destination
node of the receipt of data directed to said client.

35 106. The computer-readable storage medium for Claim 100 wherein determining physical resources includes
accessing a highly available database that stores a list of physical resources.

107. The computer-readable storage medium of Claim 106 wherein said highly available database is accessible
by said active nodes, whereby said active nodes have consistent configuration data.

100

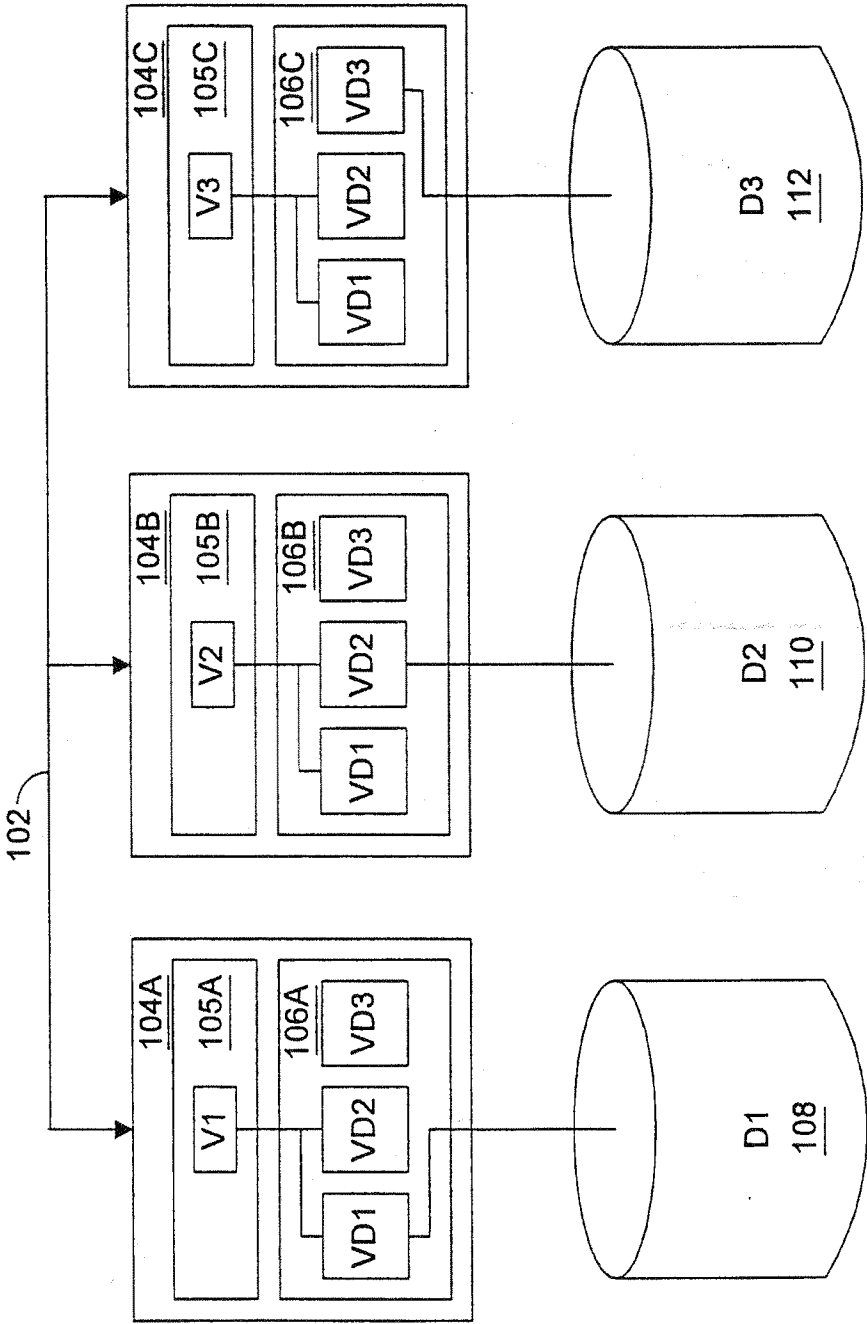


FIG. 1

200

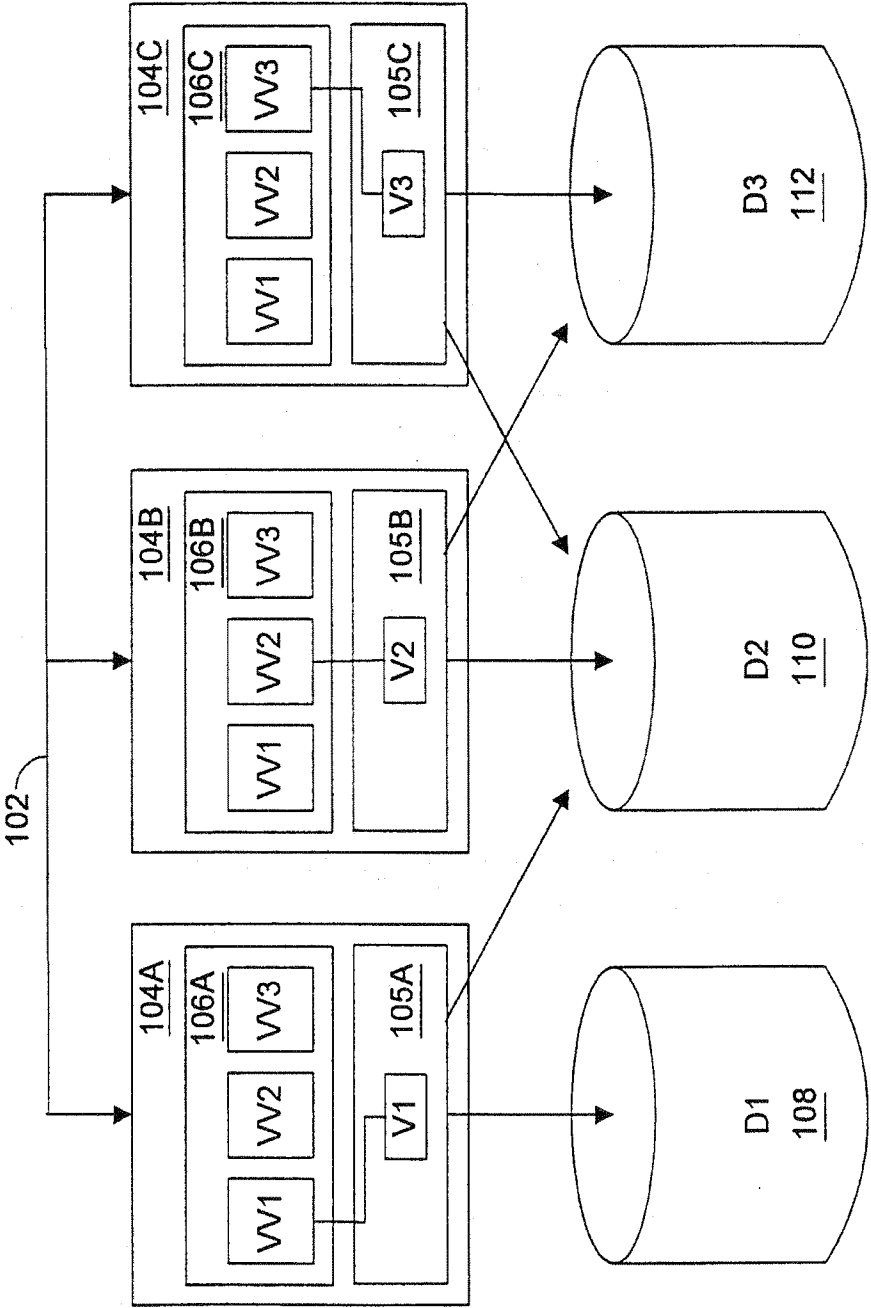


FIG. 2

3/13

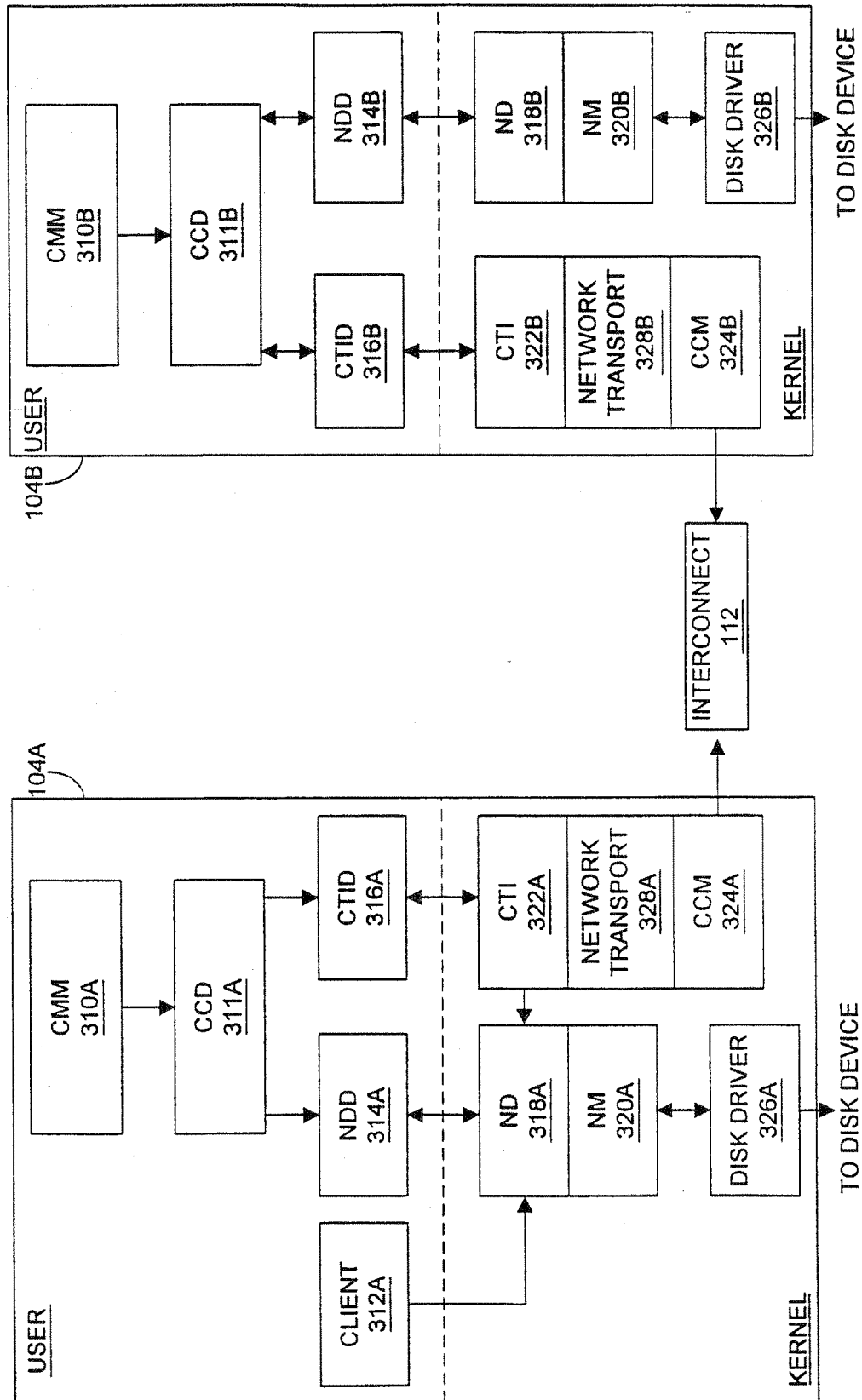


FIG. 3

4/13

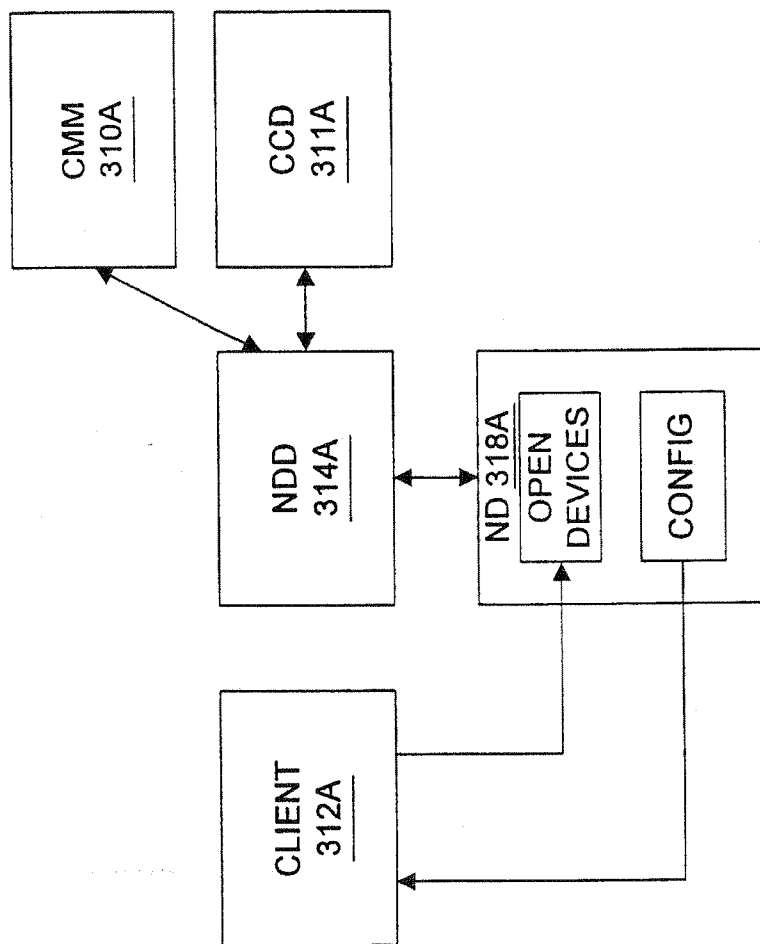


FIG. 4

5/13

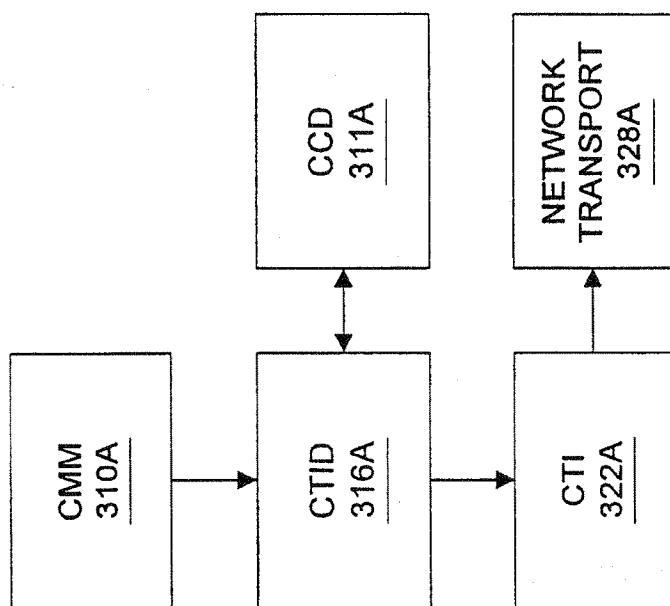


FIG. 5

6/13

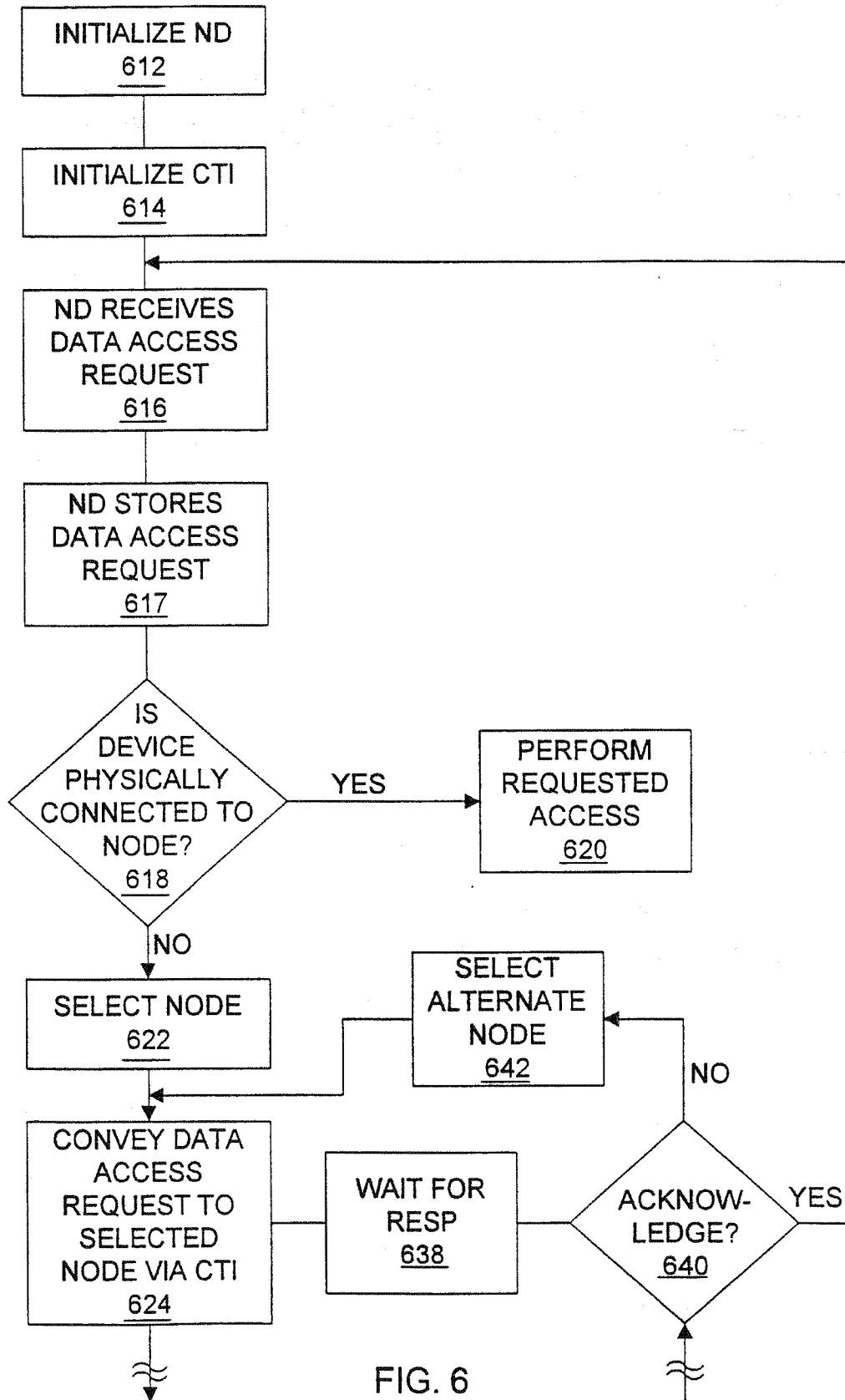


FIG. 6

7/13

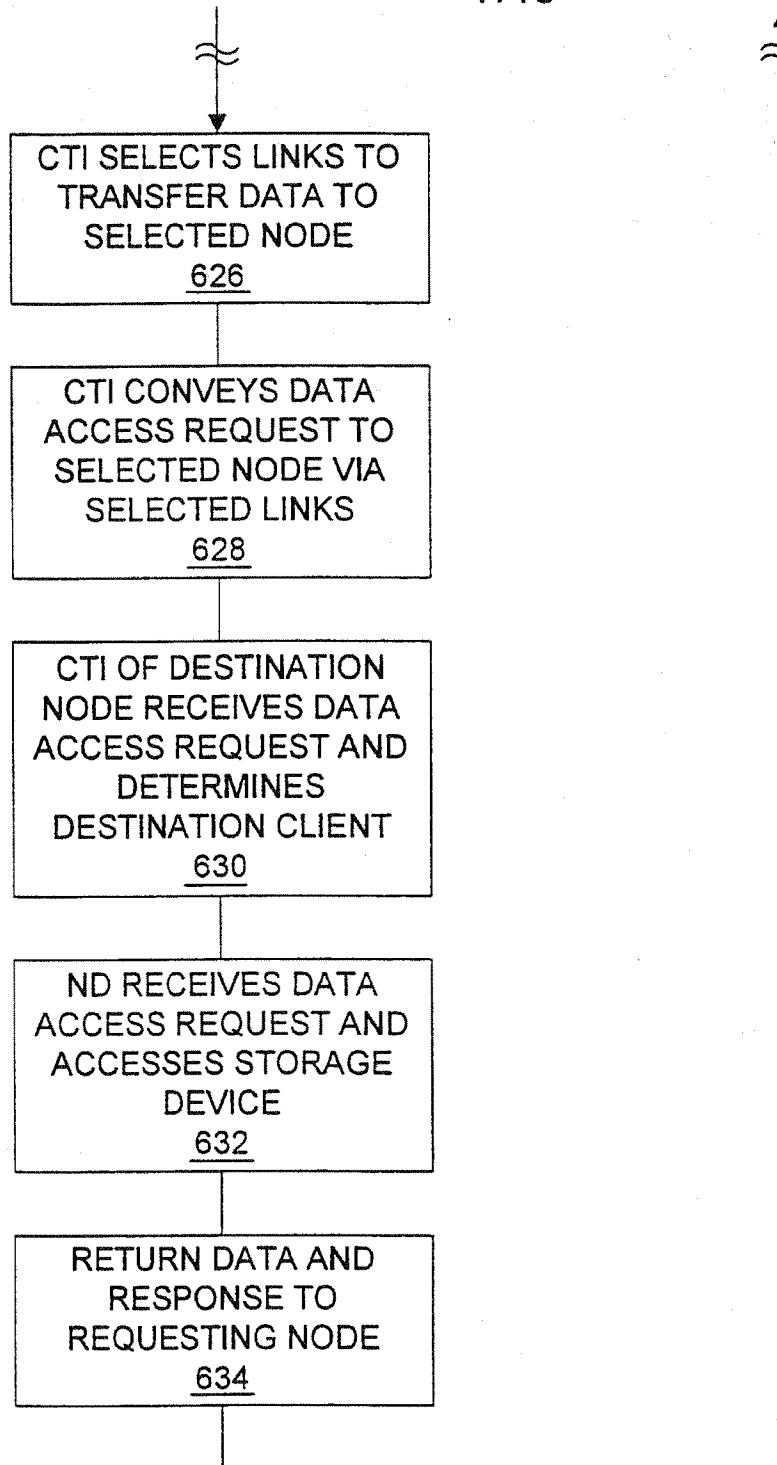


FIG. 6(CONT.)

8/13

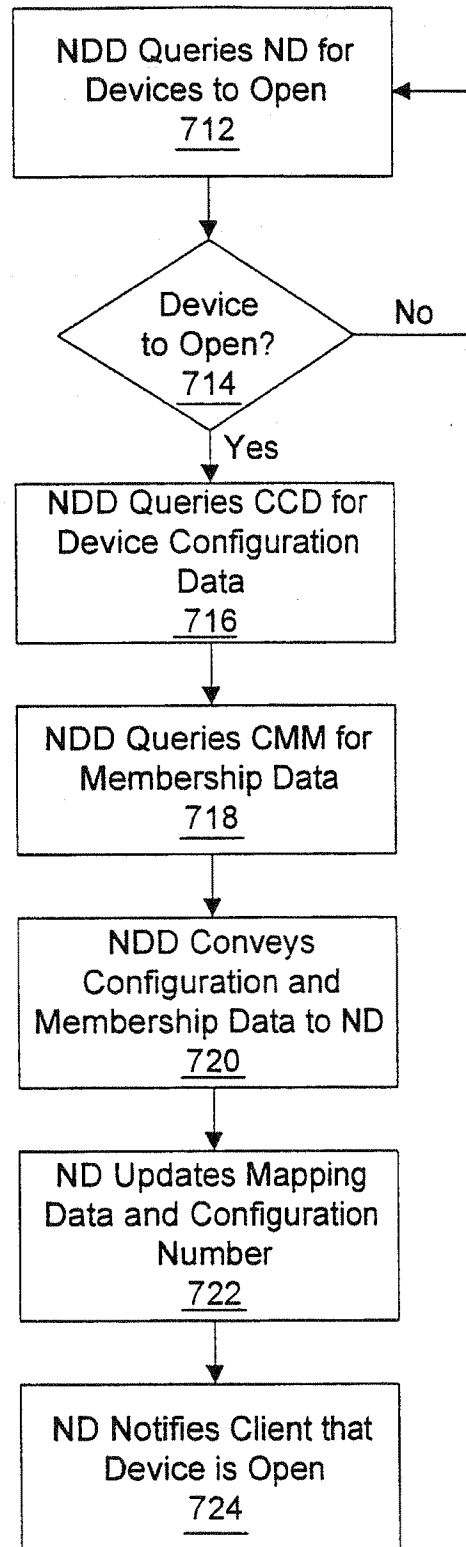
612

FIG. 7

9/13

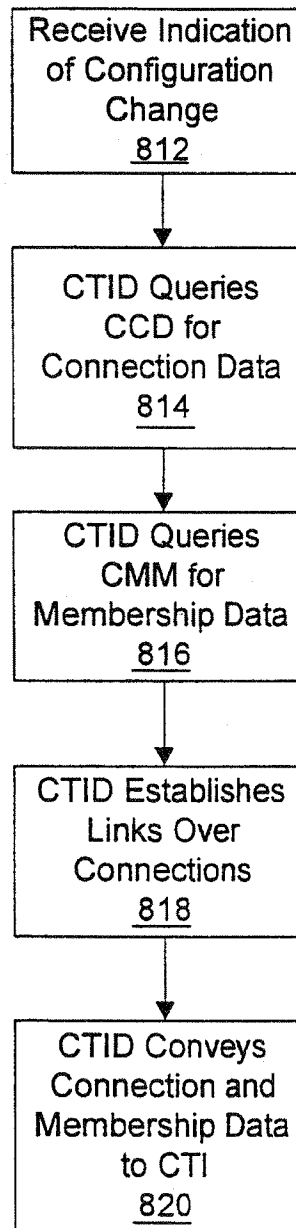
614

FIG. 8

10/13

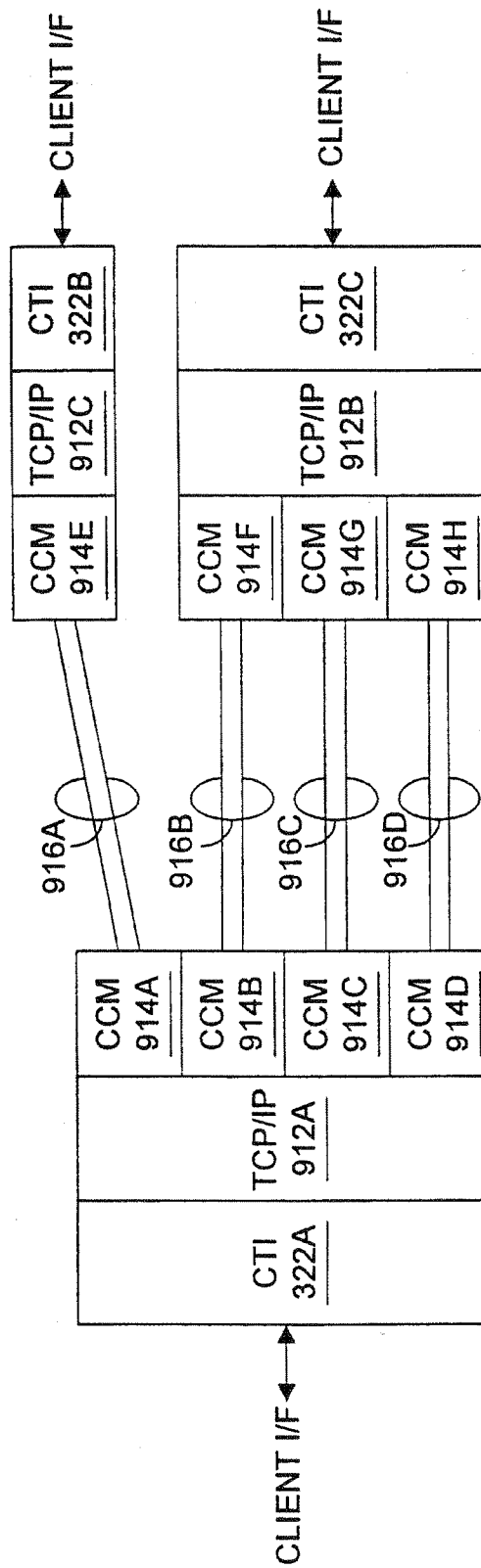


FIG. 9

11/13

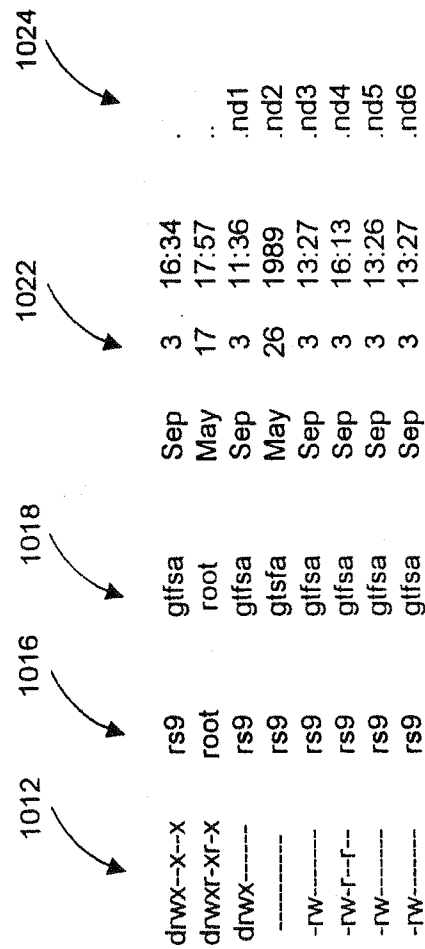


FIG. 10

12/13

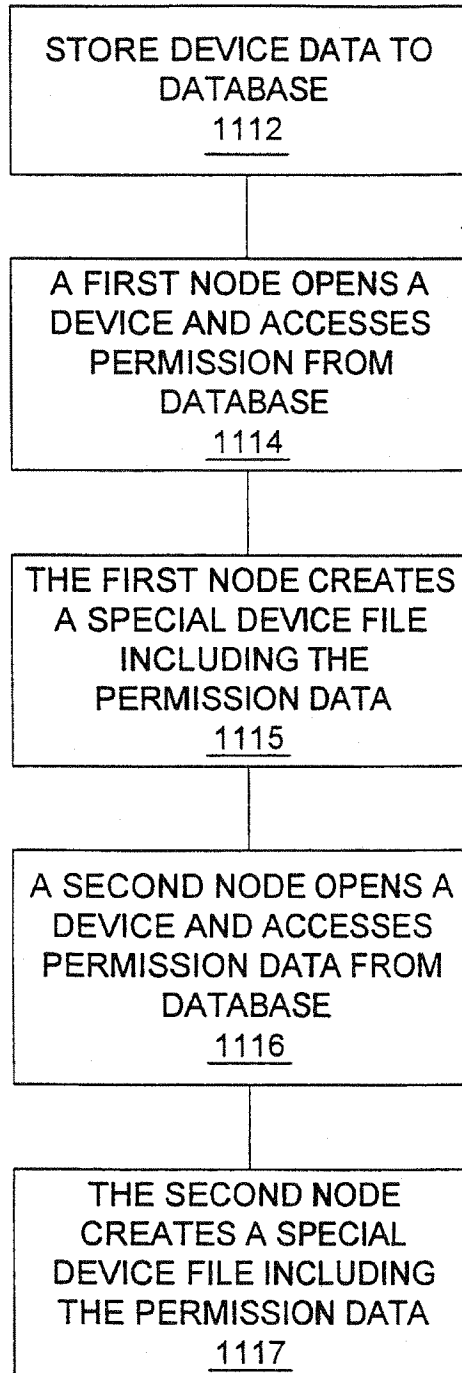


FIG. 11

13/13

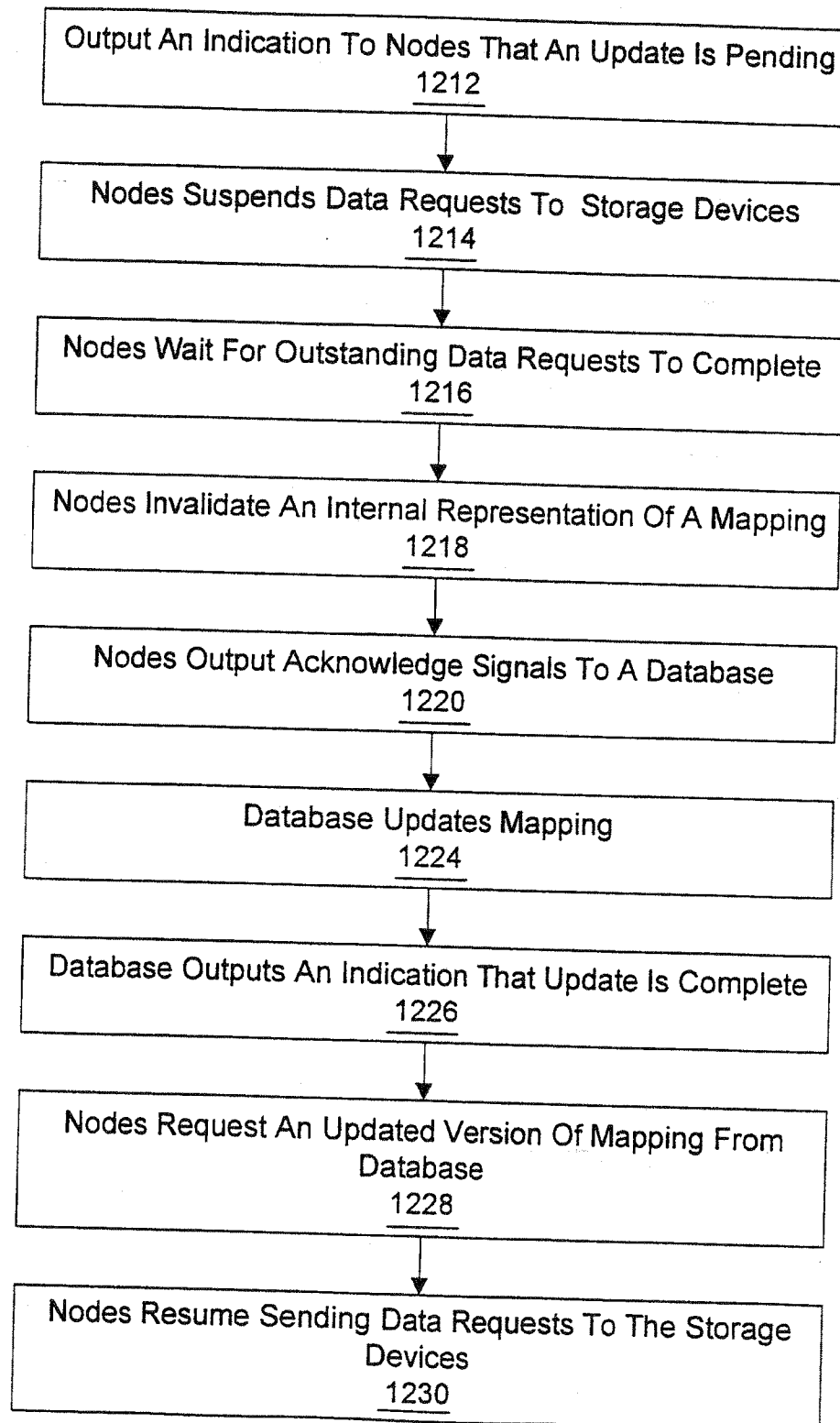


FIG. 12

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/09903

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F11/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 709 779 A (INTERNATIONAL BUSINESS MACHINES) 1 May 1996 (1996-05-01) column 5, line 18 -column 6, line 35	1-107
A	US 5 475 813 A (INTERNATIONAL BUSINESS MACHINES) 12 December 1995 (1995-12-12) abstract	1-3

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

27 September 1999

Date of mailing of the international search report

04/10/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Corremans, G

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 99/09903

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
EP 709779	A	01-05-1996	JP 8255122 A	01-10-1996
			US 5668943 A	16-09-1997
US 5475813	A	12-12-1995	NONE	